ILLINOIS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

PRODUCTION NOTE

University of Illinois at
Urbana-Champaign Library
Large-scale Digitization Project, 2007.

Technical Report No. 525

LITERACY ASSESSMENT IN A DIVERSE SOCIETY

Georgia Earnest García
P. David Pearson
University of Illinois at Urbana-Champaign

April 1991

Center for the Study of Reading

TECHNICAL REPORTS

College of Education
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
174 Children's Research Center
51 Gerty Drive
Champaign, Illinois 61820

CENTER FOR THE STUDY OF READING

Technical Report No. 525

LITERACY ASSESSMENT IN A DIVERSE SOCIETY

Georgia Earnest García
P. David Pearson
University of Illinois at Urbana-Champaign

April 1991

University of Illinois at Urbana-Champaign
51 Gerty Drive
Champaign, Illinois 61820

The work upon which this publication was based was supported in part by the Office of Educational Research and Improvement under Cooperative Agreement No. G0087-C1001-90 with the Reading Research and Education Center. The publication does not necessarily reflect the views of the agency supporting the research.

Abstract

Various formal and informal literacy assessment measures are described and evaluated. Specifically, the extent to which such measures reflect or distort the literacy performance of students from diverse linguistic, cultural, and/or economic background is noted. Although a flexible approach to assessment--giving teachers the freedom to situate or contextualize assessment--is supported, teachers in particular and educators in general are warned that they need to increase their knowledge base about language, culture, and literacy. Specific recommendations for changes in assessment practices and policies are delineated, and a list of criteria is presented for educators to use in creating and evaluating literacy assessment measures.

LITERACY ASSESSMENT IN A DIVERSE SOCIETY

Constructivist views of comprehension have dominated our thinking about reading since the early 1980s. The rhetorical trilogy of reader, text, and context is played out almost daily in education journals, state curriculum guides, basal reader philosophy statements, and methods textbooks.

Along with the constructivist view of reading has come a call for assessment measures that focus on how readers construct meaning (see Pearson & Valencia, 1987; Wixson, Peters, Weber, & Roeber, 1987). Among reading educators, process has replaced product as the primary focus for assessment, bringing into question the wide range of performance measures that have dominated the reading field for the last 40 years. Interestingly, in their pursuit of alternative measures, several reading researchers are beginning to take on an "emic" or insiders' view perspective regarding assessment (Johnston, 1989) that until recently was the purview of qualitative sociologists, educational anthropologists, and sociolinguists (Cicourel, 1974; Ogbu, 1982; Troike, 1984). This shift in orientation toward understanding how individuals within a culture construct and interpret meanings has led to the realization that all performance measures, even those with the most impeccable reputations for objectivity, are inherently interpretive; at the very least, they reflect the values, norms, and mores of the test writers who developed them and of the educators and politicians who requested or authorized them. It also has caused some educators to reject the "sorting" and "gatekeeping" functions of many of the commercially produced assessment measures (for a historical account, see Karier, 1972).

Increasingly, the rhetoric of the field calls for assessments that tell us how students approach, monitor, and process text. Critics of the conventional wisdom call for classroom-based assessment that is useful to the teachers and students involved (Goodman, Goodman, & Hood, 1989; Johnston, 1989; Resnick, 1989; Valencia, McGinley, & Pearson, 1990).

While these developments may be positive for the field of reading, new forms of assessment will not, in and of themselves, improve the education of students from diverse linguistic, cultural, and economic backgrounds. Such an improvement requires a new multicultural awareness among educators in general and reading educators in particular. They must confront the legacy of three-quarters of a century of racism and discrimination inherent in literacy assessment. They must understand how tests (and, for that matter, many forms of classroom-based assessment) have been used, albeit not always intentionally, to blame students' diversity (themselves, their families, or their communities) for their lack of growth in school-based literacy. They must understand that schools, programs, and teachers contribute to the failure many students experience in acquiring these literacy skills. Without this awareness, it is possible that new assessment measures, even those based on a constructivist view of reading, even those that "empower" teacher decision making, will hinder rather than aide students' literacy development.

Our basic thesis in this report is that the keys to meeting the assessment needs of a diverse student population are *a flexible approach to assessment* and a *dramatically improved teacher knowledge base.* We need to grant teachers greater latitude in deciding what is appropriate for a given student in a given group for a given text and a given task. In other words, teachers need the freedom to "situate" or "contextualize" assessment practices. But the minute we suggest greater freedom of choice, we are confronted with issues of accountability (Are we acting responsibly?) and equity (Are all students being treated fairly and equally?). To answer these challenges, we advocate nothing short of increased teacher knowledge about language, culture, and literacy.

The first step in developing this knowledge base is to persuade educators to consider the extent to which assessment methods distort or reflect the literacy development of students from diverse linguistic, cultural, and/or economic backgrounds. To that end, we begin this report with a quick review of the purposes of various forms of reading assessment. Then we describe some of the different assessment tasks that have been used to evaluate children's literacy development. We point out how the assessment

tasks themselves or educators' interpretations of the tasks have differentially affected and/or reflected the literacy performance of students from diverse backgrounds. Then, based on this review, we take the second important step in developing this knowledge base: We present a set of principles or guidelines that we think will be helpful in creating or evaluating the usefulness of different assessment approaches for different populations.

The Role of Assessment in Decision Making

Educators evaluate students' literacy performances for a variety of purposes. For better or worse, decisions have been made about a variety of entities, phenomena, or people. Commercially developed tests have been used to determine if programs are effective or if schools and teachers have been doing their jobs (Haney, 1985; Johnston, 1989; Pearson & Valencia, 1987; Resnick, 1989). They also have been used to direct children's placement and to document individual children's progress (Aronson & Farr, 1988; Haney, 1985; Madden & Slavin, 1987; Slavin & Madden, 1989). Standardized test scores have played a major role in determining who attends college, who is placed in college-bound tracks in our secondary schools, and who is eligible for special programs (Durán, 1983; Mercer, 1977). They have even been used, as recent history has documented, to determine who is eligible for special kindergarten programs (such as in Georgia, "Faculty Senate," 1988).

Commercially developed tests, including those found in basal reader programs, also have guided instruction (see Brophy, 1979; Dorr-Bremme & Herman, 1986). Many teachers rely on the pre- and posttests in the basal programs to determine when children are ready to progress to higher levels or to new skills (Barr & Dreeben, 1983).

One reason that commercially developed measures have had such a powerful influence on American education is that traditionally they have been viewed as "objective" and "nonbiased" (see Johnston, 1989; Stallman & Pearson, 1990). Informal measures used by teachers to help them make daily instructional decisions in the classroom have not been viewed with the same type of deference and respect as commercial tests. The reluctance of the education community to privilege informal measures is due in part to those measures' heavy reliance on teacher judgment. For some, teacher judgment is a thin disguise for subjectivity, potentially biasing the assessment process.

Teacher-made tests, a third type of assessment measure, really have not been thoroughly investigated. The limited information available suggests that these tests do not differ very much from commercially developed tests in their format and content emphases (Calfee & Hiebert, 1990). Given the abundance of commercial models available, this similarity should not be surprising.

Interestingly, as reading researchers have juxtaposed what they know about the reading process with what they see being measured on commercially developed tests, they have begun to emphasize the importance of holistic evaluations of how students approach, interpret, and engage in authentic literacy tasks (among others, see Cambourne & Turbill, 1990; Goodman et al., 1989; Johnston, 1989; Valencia et al., 1990). Unquestionably, the whole language movement, with its emphasis on classroom control of curricular decision making and empowerment for teachers and students, has propelled this movement toward these more situated assessments. Because the school environment for authentic literacy tasks is the classroom, considerable attention has been directed toward the development of ongoing assessment tasks that are part of the literate classroom environment. Some of these tasks include conferencing, dialogue or response journals, oral readings and retellings, portfolios, reader logs, and student think-alouds. One characteristic shared in the use of these situated indices is that students participate in their own evaluations by helping to select representative samples of their work.

Different Types of Assessment

To facilitate our review of assessment measures, we have chosen to focus on two types of assessment—formal and informal. Formal measures refer to those literacy tests that have been based on, or at least strongly influenced by, the standardized testing paradigm. Most of these tests are commercially produced and marketed; they include curriculum-based tests, such as those found in basal reader programs, although both standardized test publishers and basal publishers maintain that their tests serve different decision-making functions. Informal, or situated, literacy measures refer to the different types of evidence that teachers use or could collect in daily interactions with students. While teacher-made tests are another category of assessment commonly used in the classroom, we have not chosen to discuss them in our review because so little is known about them.

Formal Literacy Measures

Early reading tests. In a content analysis of current reading readiness and early reading measures, Stallman and Pearson (1990) reported that almost all of the tests reviewed measured children's performances on isolated skills in a decontextualized setting far removed from the book and print awareness features that have been emphasized so much in recent work within the emergent literacy tradition. The reading readiness tests they analyzed placed considerable emphasis on skills that many test publishers consider to be prerequisite to reading (hence, the term readiness): letter recognition, sound-symbol correspondences, oral vocabulary, key sight words, perception of shapes. Children were asked to recognize words, letters, sounds, or what they thought they heard being read. They were not asked actively to produce or identify language, nor were they asked actively to construct meaning. This was true for both standardized reading readiness tests and for basal readiness tests. The remarkable similarity between these two types of readiness tests suggests that developers of these tests work with an eye on what other developers are up to. First-grade reading tests were very similar to the readiness tests except that they focused slightly more on reading comprehension and asked children to recognize information that they had read instead of heard being read.

Clearly, a subskills approach to reading is implicit in both the reading readiness and early or first-grade reading tests. Edelsky and Harman (1988) point out that many of the skills tested, such as the ability to re-sort syllables, are not needed for reading. The emphasis on recognition tasks also means that no information is provided as to how students operationalize these tasks when they read.

Problems occur when teachers rely on tests that ask students to identify unfamiliar pictures or vocabulary or when students' prereading potential is based on their pronunciation of Standard English (Edelsky & Harman, 1988). Either due to differences in language and/or literacy experiences, children from diverse linguistic, cultural, and/or economic backgrounds frequently are placed in transitional kindergarten and first-grade programs where they are exposed to the same type of activities that are measured on the readiness tests in an attempt to get them "ready" to read (Karweit, 1989). The unfortunate consequence has been that these children are not exposed to the types of literacy activities that are thought to help promote emergent literacy and print awareness (Edwards & García, in press; Mason, 1980; Teale & Sulzby, 1986). In addition, they have been given the message that a principal goal of early reading is to be able to recognize letters, sounds, and sound-symbol correspondences (Stallman & Pearson, 1990).

Standardized reading achievement tests. Reading educators from a variety of perspectives have questioned the wisdom of overrelying on standardized reading test scores for placement and instructional purposes (Edelsky & Harman, 1988; Johnston, 1984b; Royer & Cunningham, 1981; Valencia et al., 1990). Current versions of these tests typically present students with a selection of relatively short passages reflecting a variety of genres (fiction, expository text, poetry, advertisements, and letters to the editors) followed by a series of multiple-choice questions to which there is only one correct answer. Children are asked to complete the tests within a prescribed time period, and their performances

typically are judged against the performance of other children who have taken similar versions of the tests. Test developers have tried to offset the differential influence of background knowledge by including a wide range of topics, eliminating questions that could be answered without reading the passages (passage- independent questions), and/or statistically controlling the influence of prior knowledge through latent trait theory based on population level differences (Johnston, 1984a). Nonverbal test scores frequently are included with the standardized test reports in an effort to differentiate between children's reading and reasoning abilities (Johnston, 1981).

A major problem with these tests is that they obscure rather than confront the influence of a student's prior knowledge, reading strategies, or reasoning strategies (Johnston, 1984a; Royer & Cunningham, 1981). As a result, it is difficult to know why any individual student does poorly on these tests. Other critics have pointed out that the brief and contrived test passages only simulate reading (Edelsky & Harman, 1988) and do not show what children can and cannot do with authentic literacy tasks. Furthermore, qualitative analyses of how students determine their answers to the tests have revealed that the answer selections do not always reflect the quality of students' ongoing construction of meaning and problem-solving strategies (Cicourel, 1974; García, in press; Langer, 1987). A common thread in these studies is that there are lots of "right" reasons why students select "wrong" answers.

The historically weak performance of linguistic and culturally diverse students on such tests (Durán, 1983; Mullis & Jenkins, 1990) has prompted complaints of cultural bias. Bias may occur when the test procedures and test content "reflect the dominant culture's standard of language function and shared knowledge and behavior" (Tyler & White, 1979, p. 3). *Test-wiseness*, or students' capacity to utilize characteristics and formats of the test and/or the test-taking situation to receive a high score (Millman, Bishop, & Ebel, 1965, p. 707), is a factor that may confound students' reading test performance. Most critics stress the likelihood that majority students will be more test-wise than minority students. Cicourel (1974) warns that some children "may view the task or language used as strange, yet provide a response the adult interprets as fitting the framework of the test" (p. 303).

Test developers have tried to eliminate bias by examining the concurrent or predictive validity of individual tests and by looking at the possible bias in item selection procedures, examiner characteristics, and language factors (Linn, 1983; Oakland & Matuszek, 1977). Another factor that has been investigated is *speededness*, or the failure to complete all the items on a test due to prescribed time limitations (García, in press; Mestre, 1984; Rincón, 1980). Findings from bilingual research suggest that second-language students may need more time to complete standardized test items than their monolingual counterparts because bilingual subjects tend to read slower in their second language (Kellahan & MacNamara, 1967; Mägiste, 1979).

A number of researchers have suggested that standardized tests such as the Scholastic Aptitude Test and the Graduate Record Examination do not have the same predictive validity for Hispanic and African-American students' college performance as they do for Anglo (non-Hispanic white) students (Durán, 1983; Goldman & Hewitt, 1975). In a study comparing the expository test performances of Hispanic and Anglo students at the upper primary levels, García (1988) found that the predictive validity of the students' scores on prior knowledge, vocabulary, and standardized reading tests was greater for Anglo children than it was for Hispanic children. One reason that some of the Hispanic students' expository reading test performance might have been underpredicted was that they knew less about the passage topics and test vocabulary prior to reading the test than did the Anglo students. Their lower performance on these two variables is consistent with other researchers' claims that the standardized test performance of linguistically and culturally diverse students is adversely affected by their differential knowledge of test vocabulary (Durán, 1983; Hall, Nagy, & Linn, 1984) and test topics (Royer & Cunningham, 1981).

The role of English language proficiency in second-language children's test performance is also reflected in the higher scores that they generally attain on nonverbal tests of intelligence (Durán, 1983). Yet

attempts to translate achievement tests from English to another language or to use nonverbal test scores to interpret second-language children's achievement have not always been successful. A dilemma with translated tests is that concepts do not always translate directly from one language to another (see Cabello, 1984; Durga, 1979). Likewise, juxtaposing children's standardized achievement test scores with their nonverbal test scores does not necessarily explain performance discrepancies. In one case, a computerized printout sent to the parents of a Thai child enrolled in an all-English medium school stated that, based on the child's relatively high nonverbal performance and low standardized achievement performance, the child was not working up to his potential and, therefore, needed to be encouraged to work harder. The computer program did not take into account the fact that the child was learning English as a second language, and that this situation, rather than the child's lack of effort, probably accounted for the test score discrepancy.

Domain-referenced/basal reading tests. While criterion- and domain-referenced tests may differ from norm-referenced tests in their purpose and use, they do not differ much in their format and content (Calfee & Hiebert, 1990; Stallman & Pearson, 1990). Criterion- and domain-referenced tests do not compare a child's academic performance to a representative sample or norm group. Instead, individual scores are "referenced" to some preestablished standard, with 80% typically used as the cut-off criterion. The content of such tests is based on the test developer's specification of the objectives (criteria) that the children are expected to attain, or on the knowledge that the test developer generally assumes is pertinent to a particular domain (Johnston, 1984b). These tests frequently are used by districts and teachers to set the pace of instruction or to group and place children.

One of the major problems with criterion- and domain-referenced tests turns on how they are interpreted, or on the meaning that parents or educators attribute to them. While the tests reflect the curriculum taught in basal programs, they do not always reflect how well children can comprehend text; indeed, an analysis of basal tests across grades (1-6) revealed that only about 30 to 50% of the items focused on comprehension activities (Foertsch & Pearson, 1987). So to infer that these tests are indices of reading performance is to attribute to them a status they do not merit. Secondly, because these tests are so similar to standardized reading tests in their format and content emphases, they are subject to many of the same criticisms. Finally, inherent in these tests is the assumption that the children taking them are familiar with the test content and format. According to one bilingual teacher, this has been a problem for second-language children who are "transitioned out" of bilingual classrooms into all-English classrooms. For a variety of reasons, it is not unusual for bilingual children to receive reading instruction in the bilingual classroom based on some variation of a language experience approach. Yet when these children enter the monolingual classroom, many of their teachers use basal reading tests to determine where they should be placed in the basal reading series. These series tend to assume that children have acquired a certain range of vocabulary and background knowledge. While children from the bilingual classroom may have developed the comprehension strategies needed to read, they will not necessarily have the required vocabulary or background knowledge to do well on the basal tests nor are they likely to be familiar with the type of decontextualized tasks frequently found on basal worksheets and skill tests.

New statewide reading tests. In an attempt to reflect current reading research, new statewide tests have been developed (Pearson & Valencia, 1987; Wixson et al., 1987). These tests assess children's prior knowledge of the topics, ask questions based on inferencing and text structure taxonomies, and evaluate children's awareness of reading strategies. Responding to criticisms about the brevity and lack of authenticity of the typical test passages, they have provided longer, noncontrived passages. On some of the tests, multiple answers are elicited to allow for multiple interpretations or partial interpretations of the text. Children's attitudes and interests in reading also are assessed.

Although these tests conform more to current reading comprehension theory, they still are product measures based on "mainstream" reading performance and are subject to the same complaints of bias that plague standardized tests. So far, no one has studied the extent to which they help to explain the relationship between linguistically and culturally diverse children's reading test performance and their

literacy development (García, in press). In fact, one of the potential problems with these statewide assessments is their "level playing field" mentality. Because they are usually given at specific grade levels (for example, grades 4, 8, and 12), there tends to be a common grade-level test for all students. With the push for longer passages, there is not usually a range of difficulty or topics in the passages used at a particular grade level. Obviously, then, the passages used at any given grade level will be incredibly difficult for students reading below grade level or for students who happen to be unfamiliar with the passage topics. Some students may give up, complete the task with overt hostility, or otherwise subvert the testing process; hence, inferences drawn about certain individuals are likely to be based upon measurement errors associated with extremely low or random performance. On the other hand, when individual scores are aggregated to form group scores, the likelihood of unreliable judgments decreases. As long as inferences from these assessments are limited to school or district programs, they are less prone to these criticisms and are less likely to have a negative impact on individual children. On the positive side, when scores are limited to classroom or school averages, they are more likely to focus educators' attention on aspects of the classroom or school program, rather than on individual difference variables that might be contributing to low individual performance.

Informal Literacy Measures

Since 1987, when the critiques of commercial tests intensified (and, as the emergent literacy, literature-based reading, and whole language movements gathered momentum), the field has witnessed a significant increase in the number of journal articles describing and promoting informal measures—practices that teachers can engage in daily to evaluate how their students are developing as literate individuals. In this section, we address these measures. We have included somewhat more "formal" techniques, such as oral miscue analysis and the informal reading inventory, because these measures involve teacher judgment and lend themselves to teacher adaptation and personalization.

Anecdotal records. Anecdotal records are frequently touted in the whole language literature as important indices of individual children's ongoing development (Goodman et al., 1989). For instance, a kindergarten teacher might record on a checklist when individual children read their first and/or last names, write their names, or read certain signs or logos. This type of assessment encourages teachers to focus on what children *can* do at different points in time instead of focusing on what they *cannot* do.

While anecdotal records may avoid the cultural bias implicit in many of the commercial measures that ask young children to respond to predetermined vocabulary items (for a discussion of this, see Hall et al., 1984), they are clearly dependent on the teacher's ability to recognize how individual children are responding to the classroom environment or defining the literacy tasks at hand. The quality of the data they will yield for culturally and linguistically diverse children may well depend on, among other factors, teachers' abilities to create a risk-free environment.

This point was poignantly illustrated in a public broadcasting television special, *First Things First* (WQED, 1988), in which viewers see Holly, a little girl from white Appalachia, at home and at school. At home, Holly takes the film crew around her yard, shows them her family's strawberry patch, and raves about the apples that are sweet and good to eat. She is very verbal and clearly at ease talking to the film crew about where she lives. The scene switches to her school, where she is reluctant to talk during a game, pointing instead of verbalizing. In a later scene, the teacher makes the observation that because children like Holly are allowed to grunt at home, they do not talk in school. Clearly, the teacher has not seen Holly at home. If she were to keep an anecdotal record of Holly's school-based literacy development, she might very well underestimate the girl's real competence.

Teacher-student interactions. Observing how a child interacts with an adult and/or reading materials during storybook reading is another way of assessing children's literacy development (Morrow, 1990). Morrow points out that the interactive dialogue engaged in by both the adult reader and the child reveals what the child knows about the story, how well the child understands the story, what the child

is focusing on during the story reading, and how the child is integrating background knowledge with information from the story to comprehend it.

Before teachers can use adult-child storybook interactions in multicultural classrooms, however, they need to make sure that all of the children who are participating are accustomed to this type of literacy event. Apparently, adult-child storybook reading is more common in some subcultures than it is in others (Heath, 1983; Teale, 1986). Heath's comparisons of the literacy events and adult-child interactions around these events in three subcultures in the rural South (working-class whites, working-class African-Americans, and middle-class whites) revealed that storybook interactions similar to those in school classrooms characterized the middle-class families but not the other two working-class families.

Even observing how children respond to teacher questioning in whole-group or small-group sessions may not accurately reflect what some children know about the reading task at hand. As Hymes (1972) has noted,

> It is not that a child cannot answer questions but that questions and answers are defined for [the child] in terms of one set of community norms rather than another, as to what counts as questions and answers, and as to what it means to be asked or to answer. (p. xxxi)

Philips' (1972, 1983) work with Native-American children on the Warm Springs Indian Reservation and Boggs' (1972, 1978) research with Hawaiian children indicated that these particular children were used to participant structures (teacher-student verbal interactions) that differed from those characteristic of the typical classroom. Philips warns that the reluctance of the Native-American children to participate in the structures preferred by the teachers meant that the teachers were unable to use their normal means of assessment, that is, sequencing of questions and answers, to determine the appropriate levels of instruction for the children. Philips felt that this especially hindered the teachers' implementation of reading instruction in small groups.

Story retellings. Story retellings recently have been heralded as one way for teachers not only to facilitate children's comprehension but also to assess it (Morrow, 1989). Morrow spells out several ways for teachers to use children's story retellings to assess children's reconstruction of meaning.

However, as with classroom observational data, story retellings need to be contextualized so that *all* children in the classroom are invited to participate. Leap (1982), in a microanalysis of a Native-American student's classroom behavior, found that the student barely responded when she was asked to retell a story she had read in class. On the other hand, when she was asked to make up a story about a picture drawn by a classmate, she produced an extensive narrative.

If retellings are used to assess the English reading performance of bilingual students, then teachers need to understand that some of these children may present richer protocols if they are allowed to present their retellings in their first language. Eaton (1980) discovered that Mexican-American, limited-English-proficient children were able to demonstrate longer and more accurate recalls of English text if they were allowed to produce their recalls in Spanish. Similar findings also were reported by Chamot (1980) with language-minority children and by Lee (1986) with college students learning Spanish as a foreign language. García (in press) found that bilingual children participating in all-English classrooms demonstrated greater comprehension of an English reading test than their scores indicated when she translated unknown words in the test questions into Spanish and allowed the students to code-switch or use Spanish when they explained their answers or talked about what they had read. These findings obviously pose a dilemma for the monolingual English-speaking teacher who may be working with bilingual children in a monolingual setting.

Portfolios. Sampling of student work is becoming increasingly common as a means through which teachers can determine individual students' progress and grades. In portfolio assessment, examples of

a student's writing frequently are stored in a folder, which the teacher can use to evaluate the student's written literacy development. This method differs from the type of writing assessment that increasingly is included on commercially developed tests where students write on a prescribed topic within a set time period. By using a portfolio approach, students frequently are allowed to choose their own topics, have time for planning and reflection, make revisions, and, in some instances, choose representative samples of what they consider to be their best work.

Atwell (1987) has developed a similar approach for reading. Students record their self-selected readings in a reader's log and keep a literary-response journal. This information, along with the student's goals, becomes the focus of an individual student-teacher conference held at the end of each grading period. Other types of activities that could provide the documents for a portfolio are taped oral readings and a collection of responses to reading assignments (book reports, critiques, research reports, dramatizations).

An advantage of portfolio assessment is that it allows students to display what they have learned. If artifacts are collected throughout the school year, then the progress and effort that students have made over time are revealed. For this type of assessment to work, however, students have to be motivated to perform, and drafts of their work have to be kept.

Motivation is an important issue, and crucial to the participation of low-achieving readers. Johnston and Allington (1991) point out that task-involving activities motivate the child to become involved in the activity for the sake of carrying it out, not for the sake of competing or displaying knowledge, the two prime motives in what they call ego-involving activity. If the task is viewed as ego-involving, then low achievers are likely to avoid it, will not ask for help (asking for help indicates that they cannot do it), or will set unrealistic goals for themselves. For children from diverse linguistic and cultural backgrounds, teachers have to be aware of cultural mores and norms that may influence the children's participation. It was not until the KEEP program in Hawaii adopted "talk story" (verbal interaction patterns based on Hawaiian culture) as an integral part of its reading comprehension instruction, that the Hawaiian children in the program began fully to participate (Au, 1981).

For portfolio assessment to work, the portfolio has to be more than a folder of end products; portfolios need to document the evolutionary nature of the development of a piece as well as the history of progress of the individual child. Teachers need to keep drafts of children's work. Without drafts, teachers may not see the individual progress that children have made, nor will they know where their input is needed. Drafts also may reveal some of the conflicting demands that are inherent in the literacy development of linguistically diverse students (for a discussion, see Delpit, 1988). This point is poignantly illustrated in a young African-American woman's effort to obtain a passing grade on an essay that was a major prerequisite for entrance into a required college-level rhetoric course. On her first attempt, when she was not worried about dialect features, her writing was more fluid and complex, the relationships among her ideas were clearer, and she wrote with "voice."

> *First attempt*: When I am alone, I dream about the man I want to be with. He a man that every woman wants, and every woman needs.

When she proofed her writing on the second and third attempts, she didn't seem to know what to change and, in the process of eliminating dialect features, turned to cliches and broke her thoughts down into simple sentences. Granted, the end result was dialect-free piece, but it was also choppy and voiceless.

> *Second attempt*: I daydream alot about what my knight in shining armor will be like. He has to be everything rolled all in one and nothing suppose to be wrong with him.

> *Third attempt, and the beginning of the essay she ultimately turned in*: My make-believe man is everything. He is perfect from his head down to this toes. He's handsome, romantic and intelligent.

In such a situation, a teacher who had not kept drafts would have been unaware of the struggle experienced by the student and may even have attributed the choppiness of the last draft to a lack of sophistication rather than an attempt on the part of the student to make her writing look "conventional."

Oral miscue analysis. Several different methods have been developed to assess students' oral reading. Both oral miscue analysis, also known as the reading miscue inventory (Goodman, Watson, & Burke, 1987), and running records used in Reading Recovery (Clay, 1987) attempt to document the different types of strategies that children use when they read orally. The two procedures involve recording the child's variation from the text, noting repetitions, substitutions, insertions, omissions, and self-corrections. In oral miscue analysis, the children's miscues are evaluated in terms of the extent to which they preserve syntactic and semantic consistency at the sentence and discourse levels. Children's reading comprehension also is assessed by having the children give retellings.

The authors of the newest manual on oral miscue analysis warn teachers to be careful in their analysis of oral miscues produced by dialect or second-language speakers of English (Goodman, et al., 1987). They point out that dialect features and pronunciation errors need to be evaluated in terms of semantic and discourse consistency. Such an approach assumes that teachers will be aware of these features and will know when variations such as "he be" for "he is" are consistent in meaning. Unfortunately, teachers' lack of sophistication regarding these matters is demonstrated continuously throughout the literature on the reading of linguistically diverse children (see Cazden, 1988; García & Pearson, 1990). In a study comparing the reading instruction of bilingual children in Spanish and English, Moll, Estrada, Diaz, and Lopes (1980) discovered that bilingual children who were good readers in Spanish received limited comprehension instruction in English because their English teacher was misled by the children's non-native pronunciation of English, and thought that they were not ready for English comprehension instruction.

The extent to which the assessment of oral reading is a valuable tool may also depend on children's past experiences with oral reading and the extent to which their oral reading matches their silent reading. García (1988) found that the performance of fifth- and sixth-grade Anglo children across a variety of silent and oral reading tasks was relatively consistent; whereas the performance of fifth- and sixth-grade Spanish-English bilingual children across the same tasks was inconsistent. During the miscue analysis, one of the more adept bilingual readers stopped at words she could not pronounce and waited for the researcher to provide her with the word. When the researcher told her to continue with the reading, the child skipped over the words that she could not pronounce correctly. In an interview with the researcher (R), the student (S) explained that she could understand some words even though she could not pronounce them:

R Is it easier for you to read it, or is it easier for you to pronounce it?
S Easier for me to read it cause I can't pronounce.
R You can recognize words by reading them even though you can't pronounce them?
S Uh-huh.

This student's behavior should not be too surprising given what we know about teachers' tendencies to overcorrect the oral reading of low readers and that of children who do not speak fluent standard English.

Teachers working with second-language children also need to understand how children acquire a second language. For example, Indochinese children learning English as a second language may have difficulty with gender and tense markers because these constructs are marked differently in their native languages.

Instead of considering miscues of this variety as errors, teachers need to check to see if the child understands what is being read. Children who may have difficulty producing these constructs may exhibit little or no difficulty comprehending them (see Savignon, 1983; Troike, 1969). Overemphasis on such errors ignores the developmental nature of children's second-language acquisition and shifts the emphasis of the lesson from reading to English-as-a-second-language instruction.

Informal reading inventories. These assessment measures, sometimes commercially produced and sometimes locally developed, include brief passages and vocabulary lists that are graded in difficulty. In the usual procedure, individual children orally read word lists and passages and answer comprehension questions based on the passages. The teacher starts children at a level that is comfortable for them, and they continue to read more difficult lists and passages until they reach a level where they cannot recognize 95% of the words or answer 75% of the comprehension questions (see Taylor, Harris, & Pearson, 1988). Sometimes, children's silent reading comprehension and/or listening comprehension is assessed. Retellings and probing questions also may be used to evaluate comprehension. Generally, some type of miscue analysis is employed with the oral reading.

Even though they are often produced commercially, we have listed these inventories under informal measures because some element of teacher judgment is allowed in their administration. For example, if children are accustomed to reading a passage silently before they read it orally, the teacher can incorporate this practice into the test's administration. In order to get an accurate reading of students' comprehension of "new" text, teachers also are supposed to select passages that are based on topics relatively unfamiliar to their students.

Despite their popularity, informal reading inventories are fraught with many of the problems characteristic of the formal and informal measures that we already have discussed. Implicit in them is the assumption that the passages and vocabulary selected do indeed characterize the type of reading found at a particular grade level. Further, their sequential framework implies a linear development in children's reading ability and suggests that all children have equal access to the same materials and instruction. The emphasis on oral vocabulary reading and passage decoding assumes that children cannot understand what they cannot say (a point that we already have discussed in the section on oral miscue analysis). In addition, the setting in which the informal reading inventory is administered, where an individual student interacts on a one-to-one basis with an adult, does not always guarantee a risk-free environment (for a discussion of this issue, see Labov, 1969).

A major drawback in their use with children from diverse backgrounds is that the children's reading comprehension potential may be seriously under predicted. It is likely that these children will be more adversely affected by the inventory's reliance on oral reading then will children who speak fluent standard English. They also may be less familiar with the topics and vocabulary included in the inventories than their middle-class Anglo counterparts (Bruce, Rubin, Starr, & Liebling, 1984; García, in press). Readability formulas based on word frequency counts tend to reflect more of the spoken vocabulary of middle-class students than they do of low-income students (see Bruce et al., 1984).

Finally, although informal inventories may sample aspects of children's reading, they do not reveal what students can and cannot do with authentic text in a noncontrived setting. Specifically they do not tell us how students comprehend and process both familiar and unfamiliar text, nor do they tell us how students adjust their reading according to the purpose of the task or to their interest in the topic.

Future Assessment Directions

Those who point out flaws in the conventional wisdom bear the obligation to suggest and, ultimately, to validate alternatives. Our review reveals our clear preference for a move toward situated assessment. We prefer assessment that is grounded in the local realities of schools, classrooms, teachers, and students. We prefer assessment systems in which teachers, students, parents, and the community have

a voice in deciding what is being assessed and in how it is being interpreted. Standardized and criterion-referenced tests have not provided useful information about the literacy development of students in general, but they have been particularly misleading for students from non-middle-class backgrounds. As we have pointed out, these measures do a splendid job of pointing out the obvious—that these students do not do as well as their mainstream counterparts, but they do not tell us why this occurs, nor do they tell us what these students *can do* when they are confronted with authentic literacy tasks.

Sadly, considerable evidence suggests that teachers' and administrators' reliance on conventional measures for instructional guidance and placement has resulted in a steady diet of isolated, low-level, decontextualized tasks for children of diversity. In a survey of test use in schools, Dorr-Bremme and Herman (1986) concluded that the education of low-income elementary students had been more influenced by the commercial norm- and criterion-referenced tests used to meet federal and state program evaluation requirements than had the education of other students. Although they report that part of this was in response to the general public's negative reaction to these students' traditionally low test scores and to the large numbers of low-income children who participate in specialized programs receiving state and federal funding, they note that a good part of this reliance was motivated by principals' and teachers' concerns about low-income students' basic skills development, and their belief that commercial tests could assess this development. The end result is that teachers of low-income students reported spending more time and resources teaching their students the material on the tests than did teachers of other students.

These facts and conclusions lead us to several recommendations about changes in assessment practices and policies.

1. We (meaning the educational community at large) should reduce our reliance on group testing as indices of individual, school, district, or state accountability. Exactly how we reached our current state of excessive reliance on indirect measures of reading is not altogether clear. Surely when our incessant quest for efficiency was wedded to our desire to instill in everyone a greater sense of both personal and communal responsibility, we were led to hold students, teachers, and administrators accountable to measures that took as little time as possible to administer and score. School officials understandably want to look good on these high-stakes measures (especially when scores get published in the local media or when students can fail to advance or graduate); hence they resort to one of the oldest traditions in education—teaching to the test. This insidious practice asks a test that was probably designed to serve as a simple, indirect index of progress on a phenomenon, like reading, to serve the role of an implicit, or sometimes an explicit, blueprint for a curriculum. And the simple fact is that tests are not up to that strain. Because tests are based on a mainstream viewpoint, reflecting the knowledge and values of the majority, they tend to obscure the performance of students from diverse backgrounds.

2. We should privilege assessment traditions that are grounded in classrooms and schools. If assessment tools are going to be used to serve this high-stakes accountability function (and, by implication the curriculum blueprint function), then we need very different assessments. If assessment is going to drive school improvement, then we need methods reminiscent of responsive evaluation approaches (Stake, 1976; see also Cambourne & Turbill, 1990). Such approaches would likely involve site visitations by independent reviewers at different intervals throughout the school year. In the process of helping the school staff, students, and community evaluate the school, these reviewers would visit classrooms, talk to participants (students, teachers, administrators, parents, and community leaders), and examine a wide range of artifacts that, taken together, tell the story of literacy performance, instruction, and use within the school and the community. The approach we are suggesting is also similar to portfolio assessment systems that have been discussed recently by several literacy researchers (Valencia et al., 1990) and school reformers (Wiggins, 1989). Local involvement could be guaranteed by insisting that school participants (e.g., administrators, teachers, and students) select many, if not most, of the

portfolio entries. This type of assessment would allow schools and the communities they serve to reflect the values and competencies they deem important.

Although we have concerns about situated assessment, we think that it should dominate teachers' attempts to evaluate and nurture children's literacy development. Anecdotal records, oral readings, portfolios, story-retellings, and teacher-student interactions all provide useful information—windows into children's literacy progress in performance and disposition. These methods allow us to see what students can and cannot do across different tasks and at different points in time. This change in focus would give us a different and more informative view of what students who typically score low on formal measures are capable of doing.

With situated assessment, teachers also can support students' efforts as they provide important feedback. Documenting the extent to which students perform literacy tasks or utilize different types of strategies to construct meaning with and without "scaffolding" from teachers or peers can help inform teacher decision-making at the same time that it starts students on the road to independent self-evaluation. Self-evaluation is an important goal if we want students to be active literacy participants. If students are to construct meaning both in and out of the school setting, then they have to be able to monitor their own reading and writing without the assistance of a teacher.

3. We should take steps to ensure that teachers and administrators are knowledgeable about issues of language and culture. For situated assessment to work with students from diverse backgrounds, administrators and teachers need to take on an "emic" perspective. They need to become concerned about how students interpret (if you prefer, how they "read") the classroom context and the literacy events before them. This requires considerable knowledge and effort on the part of administrators and teachers.

First of all, educators need to know more about the influence of language and culture on children's learning. They need to understand that it is not language and cultural differences in themselves that cause learning difficulties. Rather, it is educators' misinterpretation of language and culture, as reflected in misguided remedial practices or unwarranted conclusions about children's motivation and behavior. Clearly, teacher education, both preservice and inservice, is the primary means available to the profession for helping educators to acquire this knowledge.

Second, lines of communication have to be opened to parents and other community leaders. Parents have to feel comfortable in the school context and know that it is acceptable for them, for example, to tell a teacher that they don't understand why their child is viewed as verbally unresponsive in school when the child constantly talks at home. Teachers, in turn, have to be willing to accept parents' observations. Perhaps, most importantly, teachers have to expand the range of explanations they consider as they try to understand why some children are not performing well in the classroom. They have to be willing to seek out answers, not just by sharing information with their middle-class colleagues, but also by sharing information with participants in the child's culture (see Moll, 1990). The latter may be one of the most difficult tasks at hand because it requires teachers to acknowledge that their own behavior is conditioned by their own socialization and that there are alternative ways of learning, interacting, and behaving.

4. We should promote new criteria for evaluating assessment tools. Traditionally, four criteria have dominated our evaluation of tests and other assessment devices: reliability, validity, objectivity, and efficiency. Reliability, which indexes the consistency with which an assessment device measures whatever it measures, and validity, which indexes the degree to which a tool is a true measure of what it purports to measure, have served as the cornerstones of measurement in American education. Objectivity (Is the test fair, unbiased, and independent of the views or whims of the test creators or administrators?) and efficiency (Is this the least expensive and intrusive index to be found?) have been only slightly less important in our selection of assessment tools. In fact, one of the reasons that situated assessments are often discredited hinges on their perceived lack of objectivity. Because such assessments rely on the

interpretation of individual teachers, they are viewed as being rife with opportunities for bias and even ethnic discrimination. Hence, one often hears the plea for a "level playing field" in which all students, regardless of background, have an equal opportunity to succeed. In most cases, the level field turns out to be another multiple choice standardized test, which, as we have suggested, tends to represent the values and viewpoints of the individuals who wrote the test items and the advocators and politicians who requested them.

Efficiency also exercises considerable muscle in a day and age when schools are literally inundated by assessment requirements from state mandates, federal funding requirements, and local accountability plans. It often comes in the guise of an appeal for more emphasis on instruction: "We have to minimize the time we take away from instruction, so give us a simple, quick and dirty, uncomplicated, and unintrusive test." Such a demand promotes the continued dominance of multiple-choice, standardized tests.

If we are to reduce our reliance on these four traditional criteria, what criteria will we use as alternatives? In fact, we really do not want to suggest that reliability and validity should be diminished in importance. Unreliable measures just cannot be trusted, especially when it comes to making entry and exit decisions for individuals. Without validity, a test could measure something other than what it purports to measure, and its use could be downright harmful for decision-making purposes. It may only be a slight semantic variation, but we like the meaning conveyed by the term, *trustworthiness*, a term we have borrowed from qualitative evaluation (see Guba, 1981). In a sense, trustworthiness encompasses both reliability and validity. We could not *trust* a decision based on an unreliable or invalid assessment tool. How trustworthy a test is, however, should be judged by how well it depicts a child's performance. Educators should be wary of using instruments that reveal little beyond the existence of a low score.

A second criterion which we would like to see applied to assessment tools is authenticity. Authenticity is more than face validity (Does the task look like a reading task?) or curricular validity (Is the task consistent with the manner in which it is presented in the current curriculum?). A literacy assessment task is authentic to the degree that it resembles the way literacy is used in real-life. It is not enough to be consistent with the curriculum, which itself may be disconnected with real-life literacy. A slightly less rigorous version of the authenticity criterion would be this: An assessment task is authentic to the degree that it promotes the use of literacy in the ways we expect students to really use it in genuine communication acts.

A derivative (perhaps it is a slight variation) of authenticity is what some have called *instructional validity* (Pearson & Valencia, 1987). Instructional validity is almost the logical complement of curricular validity. Recall that an assessment task is valid to the degree that it resembles the form and manner in which it is used within the curriculum. An assessment task is instructionally valid to the degree that it promotes instruction that is known to lead to student mastery over authentic literacy performance (i.e., in genuine acts of communication). In a sense, instructional validity requires us to reconsider our traditional approach to instructional research. Normally, in instructional research, we assume the validity of a test and then proceed to evaluate the validity of competing instructional approaches. We are suggesting just the opposite: We should assume (on the basis of cumulative experience) the validity of the instruction and evaluate the validity of the assessment task (by measuring the degree to which it is sensitive to assumed growth in the target behavior). This change in direction would allow us to look at what students from diverse backgrounds can do in a supportive classroom environment, and then see if the assessment measures reveal this performance.

5. We should change conventional assessments. While we remain firm in our conviction that situated assessments should dominate our instructional decision-making, we recognize the fact that "big" (i.e., wide-scale) assessment is likely to remain a part of our educational system for the foreseeable future. Thus, we need to improve this tool, however flawed in basic conception and purpose we may think it is. We applaud the efforts in Michigan (Wixson et al., 1987), Illinois (Pearson & Valencia, 1987), and

many other states to reform wide-scale assessments. What is significant about these efforts is their use of "authentic" texts and more "authentic" tasks; they are more firmly grounded in a constructive model of the reading process. As tools for program evaluation these new assessments are likely to promote exciting, alternative literacy curriculum reforms for mainstream students. We have seen changes in recent editions of commercially available standardized tests, and even the SAT is being revised to include longer passages, more thoughtful questions, and more instructionally valid tasks (Fiske, 1989). On the other hand, we have to admit that we still do not know whether these sorts of changes will result in more valid assessments for diverse populations. Due to the "level playing field" mentality of these tests, they may not be as trustworthy or instructionally valid for students from diverse linguistic, cultural, and/or economic backgrounds.

6. We should take a more realistic perspective on what assessment tasks, especially commercially available tests, can do for us. Sometimes we act as though tests were magic or divine in origin. We would be better off if we remembered a few simple facts and rules of thumb about how to use tests.

- *All tests are samples of performance.* Because tests are samples, we need to admit that they never capture the range of texts and situations to which we expect the behavior to apply.

- *Tests are surrogates for the real thing, not the thing itself.* A test is nothing but an indirect index of progress on a phenomenon we happen to care about and for which our resources for evaluation are limited. In fact it is the limited resources that force us to use tests (indirect samples) rather than direct observation of the thing itself. If we remembered this fact, we might escape the utter idiocy involved in teaching directly to tests and return to teaching to *thoughtful conceptualizations of curriculum.*

- *Multiple indices are both desirable and necessary.* Anyone would be a fool to rely on a single measure of anything that mattered to them. What we want are converging indices of progress (or the lack of progress) so that we can place greater trust in our decisions. Converging indices are especially important for students from diverse populations, for whom some measures may not be as trustworthy or instructionally valid as others.

- *Subjectivity can never be avoided, only masked.* One of the great illusions of standardized assessments is that they are more objective, and hence more trustworthy, than assessments for which teachers have to make interpretations and judgements. Yet, someone has to decide what passages to use, what questions to ask, what choices to provide for multiple-choice items, what the "right" answers are. Those someones are people who are subject to the same biases as those who make judgements in classrooms. And even if we granted, for the sake of argument, that tests could be developed in a non-subjective manner, someone would still have to decide what a score meant, and that decision inevitably requires interpretation. We would all be better off if we admitted that judgement is an inherent part of any assessment activity and, in the very next breath, suggested that the best and only guarantee against poor judgment is greater professional knowledge of literacy processes, instruction, and assessment.

These are the future directions that we would like to see assessment take. In developing and evaluating new literacy assessment methods or procedures, we think that it is especially important to keep in mind what we know about the reading/writing process, the test performance of students from diverse backgrounds, and the potential pitfalls that all assessments pose for these students. To meet these problems, we must concern ourselves with the development of teachers' and administrators' knowledge base. As a summary, we close with a list of criteria we would hope that educators use in evaluating and creating literacy assessments. Good assessments

- engage students in authentic literacy tasks

- reflect a constructivist view of reading and writing
- reveal student progress over time
- emphasize what students can and cannot do (with and without help from the teacher, other adults, or their peers)
- take advantage of rather than ignore or, even worse, penalize students' diversity
- provide multiple indices of students' literacy development and interests
- acknowledge students' interpretations (i.e., their "readings") of literacy tasks
- encourage the involvement of students, parents, and community participants

Without these characteristics, "new" methods of assessment will be no more useful to a diverse society than the "old" methods.

References

Aronson, E., & Farr, R. (1988). Issues in assessment. *Journal of Reading, 32*, 174-177.

Atwell, N. (1987). *In the middle: Writing, reading, and learning with adolescents*. Portsmouth, NH: Heinemann.

Au, K. H. (1981). Participation structures in a reading lesson with Hawaiian children: Analysis of a culturally appropriate instructional event. *Anthropology and Education Quarterly, 11*, 91-115.

Barr, R., & Dreeben, R., with Wiratchai, N. (1983). *How schools work*. Chicago: University of Chicago Press.

Boggs, S. T. (1972). The meaning of questions and narratives to Hawaiian children. In C. B. Cazden, V. P. John, & D. Hymes (Eds.), *Functions of language in the classroom* (pp. 299-327). New York: Teachers College Press.

Boggs, S. T. (1978, December). The development of verbal disputing in part-Hawaiian children. *Language in Society, 7*, 325-344.

Brophy, J. (1979). Teacher behavior and student learning. *Educational Leadership, 37*, 33-38.

Bruce, B., Rubin, A., Starr, K., & Liebling, C. (1984). Sociocultural differences in oral vocabulary and reading material. In W. S. Hall, W. E. Nagy, & R. Linn (Eds.), *Spoken words: Effects of situation and social group on oral word use and frequency* (pp. 466-480). Hillsdale, NJ: Erlbaum.

Cabello, B. (1984). Cultural interface in reading comprehension: An alternative explanation. *Bilingual Review, 2*, 12-20.

Calfee, R., & Hiebert, E. (1990). Classroom assessment of reading. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *The handbook of reading research* (Vol. 2, pp. 281-309). New York: Longman.

Cambourne, B., & Turbill, J. (1990). Assessment in whole-language classrooms: Theory into practice. *Elementary School Journal, 90*, 337-349.

Cazden, C. B. (1988). *Classroom discourse: The language of teaching and learning*. Portsmouth, NH: Heinemann.

Chamot, A. U. (1980, November). Recent research on second-language reading. *NCBE FORUM* (pp. 3-4).

Cicourel, A. (1974). Some basic theoretical issues in the assessment of the child's performance in testing and classroom settings. In A. Cicourel, K. H. Jennings, S. H. M. Jennings, K. C. W. Leiter, R. Mackay, H. Mehan, & D. Roth (Eds.), *Language use and school performance* (pp. 300-351). New York: Academic Press.

Clay, M. M. (1987). *The early detection of reading difficulties* (3rd ed.). Auckland, New Zealand: Heinemann.

Delpit, L. D. (1988). The silenced dialogue: Power and pedagogy in educating other people's children. *Harvard Educational Review, 58*, 280-298.

Dorr-Bremme, D. W., & Herman, J. L. (1986). *Assessing student achievement: A profile of classroom practices*. Los Angeles: University of California, Center for the Study of Evaluation.

Durán, R. P. (1983). *Hispanics' education and background: Predictors of college achievement*. New York: College Entrance Examination Board.

Durga, R. (1979). Memory organization, bilingualism, and interlingual interference: A comparative analysis of the semantic distance and semantic judgment of English monolingual and Spanish-English bilingual students. In *Outstanding dissertations in bilingual education* (pp. 15-28). Rosslyn, VA: National Clearinghouse for Bilingual Resources.

Eaton, A. J. (1980). A psycholinguistic analysis of the oral reading miscues of selected field-dependent and field-independent native Spanish-speaking, Mexican-American first-grade children. In *Outstanding dissertations in bilingual education* (pp. 71-86). Rosslyn, VA: National Clearinghouse for Bilingual Resources.

Edelsky, C., & Harman, S. (1988). One more critique of reading tests—With two differences. *English Education 20*, 157-171.

Edwards, P. A., & García, G. E. (in press). Parental involvement in mainstream schools: An issue of equity. In M. Foster & S. S. Goldberg (Eds.), *Readings on Equal Education* (Vol. 2).

Faculty Senate position paper on the Georgia kindergarten testing/first grade readiness policy. (1988). Unpublished manuscript, University of Georgia, Athens.

Fiske, E. B. (1989, January 3). Changes planned in entrance tests used by colleges. *New York Times*, pp. 1, 16.

Foertsch, M., & Pearson, P. D. (1987, December). *Reading assessment in basal reading series and standardized tests*. Paper presented at the annual meeting of the National Reading Conference, St. Petersburg, FL.

García, G. E. (1988). *Factors influencing the English reading test performance of Spanish-English bilingual children*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

García, G. E. (in press). Factors influencing the English reading test performance of Spanish-speaking Hispanic children. *Reading Research Quarterly*.

García, G. E., & Pearson, P. D. (1990). *Modifying reading instruction to maximize its effectiveness for all students* (Tech. Rep. No. 489). Urbana-Champaign: University of Illinois, Center for the Study of Reading.

Goldman, R. E., & Hewitt, B. (1975). An investigation of test bias for Mexican American college students. *Journal of Educational Measurement, 12*, 187-196.

Goodman, K. S., Goodman, Y. M., & Hood, W. J. (1989). *The whole language evaluation book*. Portsmouth, NH: Heinemann.

Goodman, Y. M., Watson, D. J., & Burke, C. L. (1987). *Reading miscue inventory: Alternative procedures*. New York: Richard C. Owen.

Guba, E. G. (1981). Criteria for assessing the trustworthiness of naturalistic inquiries. *Educational Communication and Technology Journal, 29*, 75-92.

Hall, W. S., Nagy, W. E., & Linn, R. (1984). *Spoken words: Effects of situation and social group on oral word usage and frequency*. Hillsdale, NJ: Erlbaum.

Haney, W. (1985). Making testing more educational. *Educational Leadership, 47*, 4-12.

Heath, S. B. (1983). *Ways with words: Language, life, and work in communities and classrooms*. Cambridge, MA: Cambridge University Press.

Hymes, D. (1972). Introduction. In C. B. Cazden, V. P. John, & D. Hymes (Eds.), *Functions of language in the classroom* (pp. xi-lvii). New York: Teachers College Press.

Johnston, P. (1981). *Prior knowledge and reading comprehension test bias*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

Johnston, P. (1984a). Prior knowledge and reading comprehension test bias. *Reading Research Quarterly, 19*, 219-239.

Johnston, P. (1984b). *Reading comprehension assessment: A cognitive basis*. Newark, DE: International Reading Association.

Johnston, P. (1989). Constructive evaluation and the improvement of teaching and learning. *Teachers College Record, 90*, 509-528.

Johnston, P., & Allington, R. (1991). Remediation. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 984-1012). New York: Longman.

Karier, C. J. (1972). Testing for order and control in the corporate liberal state. *Educational Theory, 22*, 159-180.

Karweit, N. L. (1989). Effective kindergarten programs and practices for students at risk. In R. E. Slavin, N. L. Karweit, & N. A. Madden (Eds.), *Effective programs for students at risk* (pp. 103-142). Boston: Allyn & Bacon.

Kellahan, T., & MacNamara, J. (1967). Reading in a second language. In M. D. Jenkinson (Ed.), *Reading instruction: An international forum* (pp. 231-240). Newark, DE: International Reading Association.

Labov, W. (1969). The logic of nonstandard English. In R. D. Abrahams & R. C. Troike (Eds.), *Language and cultural diversity in American education* (pp. 225-261). Englewood Cliffs, NJ: Prentice Hall.

Langer, J. A. (1987). The construction of meaning and the assessment of comprehension: An analysis of reader performance on standardized test items. In R. O. Freedle & R. P. Durán (Eds.), *Cognitive and linguistic analyses of test performance* (Vol. XXII in R. O. Freedle, Ed., *Advances in discourse processes*) (pp. 225-244). Norwood, NJ: Ablex.

Leap, W. L. (1982). The study of Indian English in the U.S. Southwest: Retrospect and prospect. In F. Barken, E. A. Brandt, & J. Ornstein-Galicia (Eds.), *Bilingualism and language contact: Spanish, English and Native American languages* (pp. 101-119). New York: Teachers College Press.

Lee, J. F. (1986). On the use of the recall task to measure L2 reading comprehension. *Studies in Second Language Acquisition, 8*, 201-211.

Linn, R. L. (1983). Predictive bias as an artifact of selection procedures. In H. Wainer & S. Messick (Eds.), *Principles of modern psychological measurement: A festschrift for Frederic M. Lord* (pp. 27-40). Hillsdale, NJ: Erlbaum.

Madden, N. A., & Slavin, R. E. (1987, April). *Effective pull-out programs for students at risk.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Mägiste, E. (1979). The competing language systems of the multilingual: A developmental study of decoding and encoding processes. *Journal of Verbal Learning and Verbal Behavior, 18,* 79-89.

Mason, J. (1980). When do children begin to read: An exploration of four-year-old children's letter and word reading competencies. *Reading Research Quarterly, 15,* 203-227.

Mercer, J. R. (1977). Identifying the gifted Chicano child. In J. L. Martinez (Ed.), *Chicano psychology* (pp. 155-173). New York: Academic Press.

Mestre, J. P. (1984). The problem with problems: Hispanic students and math. *Bilingual Journal, 32,* 15-19.

Millman, J., Bishop, C. H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement, 25,* 707-726.

Moll, L. C. (1990, February). *Literacy research in community and classrooms: A sociocultural approach.* Paper presented at the meeting on Multi-disciplinary Perspectives on Research Methodology in Language Arts, National Conference on Research in English, Chicago, IL.

Moll, L. C., Estrada, E., Diaz, E., & Lopes, L. M. (1980). The organization of bilingual lessons: Implications for schooling. *Quarterly Newsletter of the Laboratory of Comparative Human Cognition, 2,* 53-58.

Morrow, L. M. (1989). Using story retelling to develop comprehension. In K. D. Muth (Ed.), *Children's comprehension of text: Research into practice* (pp. 37-58). Newark DE: International Reading Association.

Morrow, L. M. (1990). Assessing children's understanding of story through their construction and recognition of narrative. In L. M. Morrow & J. K. Smith (Eds.), *Assessment for instruction in early literacy* (pp. 110-134). Englewood Cliffs, NJ: Prentice Hall.

Mullis, I. V. S., & Jenkins, L. B. (1990). *The reading report card, 1971-88: Trends from the nation's report card.* Princeton, NJ: National Assessment of Educational Progress, Educational Testing Service.

Oakland, T., & Matuszek, P. (1977). Using tests in nondiscriminatory assessment. In T. Oakland (Ed.), *Psychological and educational assessment of minority children* (pp. 52-69). New York: Brunner/Mazel.

Ogbu, J. (1982). Cultural discontinuities and schooling. *Anthropology and Education Quarterly, 13,* 291-307.

Pearson, P. D., & Valencia, S. (1987). Assessment, accountability, and professional prerogative. In J. E. Readence & R. S. Baldwin (Eds.), *Research in literacy: Merging perspectives: Thirty-sixth yearbook of the National Reading Conference* (pp. 3-16). Rochester, NY: National Reading Conference.

Philips, S. (1972). Participant structures and communicative competence: Warm Springs children in community and classroom. In C. B. Cazden, V. P. John, & D. Hymes (Eds.), *Functions of language in the classroom* (pp. 370-394). New York: Teachers College Press.

Philips, S. U. (1983). *The invisible culture: Communication in classroom and community on the Warm Springs Indian Reservation*. New York: Longman.

Resnick, L. B. (1989, October). *Tests as standards of achievement in schools*. Prepared for the Educational Testing Service Conference, The Uses of Standardized Tests in American Education, New York.

Rincón, E. (1980). Test speededness, text anxiety, and test performance: A comparison of Mexican American and Anglo American high school juniors (Doctoral dissertation, University of Texas at Austin, 1979). *Dissertation Abstracts International, 40,* 5772A.

Royer, J. M., & Cunningham, D. J. (1981). On the theory and measurement of reading comprehension. *Contemporary Educational Psychology, 6,* 187-216.

Savignon, S. J. (1983). *Communicative competence: Theory and classroom practice: Texts and contexts in second language learning*. Reading, MA: Addison-Wesley.

Slavin, R. E. (1987). Ability grouping and student achievement in elementary schools: A best-evidence synthesis. *Review of Education Research, 57,* 347-350.

Slavin, R. E., & Madden, N. A. (1989). Effective classroom programs for students at risk. In R. E. Slavin, N. L. Karweit, & N. A. Madden (Eds.), *Effective programs for students at risk* (pp. 23-51). Boston: Allyn & Bacon.

Stake, R. E. (1976). A theoretical statement of responsive evaluation. *Studies in Educational Evaluation, 2,* 19-22.

Stallman, A. C., & Pearson, P. D. (1990). Formal measures of early literacy. In L. M. Morrow & J. K. Smith (Eds.), *Assessment for instruction in early literacy* (pp. 7-44). Englewood Cliffs, NJ: Prentice Hall.

Taylor, B., Harris, L. A., & Pearson, P. D. (1988). *Reading difficulties: Instruction and assessment*. New York: Random House.

Teale, W. H. (1986). Home background and young children's literacy development. In W. H. Teale & E. Sulzby (Eds.), *Emergent literacy: Writing and reading* (pp. 173-206). Norwood, NJ: Ablex.

Teale, W. H., & Sulzby, E. (1986). Emergent literacy as a perspective for examining how young children become writers and readers. In W. H. Teale & E. Sulzby (Eds.), *Emergent literacy: Writing and reading* (pp. vii-xxv). Norwood, NJ: Ablex.

Troike, R. C. (1969). Receptive competence, productive competence, and performance. In J. E. Alatis (Ed.), *Monograph series on language and linguistics, 22,* 63-73.

Troike, R. C. (1984). SCALP: Social and cultural aspects of language proficiency. In C. Rivera (Ed.), *Language proficiency and academic achievement* (pp. 44-54). Avon, England: Multi-Lingual Matters, Ltd.

Tyler, R. W., & White, S. H. (1979). Chairmen's report. In National Institute of Education, *Testing, teaching, and learning: Report of a conference on research on testing* (pp. 3-32). U.S. Department of Health, Education, and Welfare. Washington, DC: Government Printing Office.

Valencia, S., McGinley, W., & Pearson, P. D. (1990). Assessing reading and writing: Building a more complete picture. In G. Duffy (Ed.), *Reading in the middle school* (2nd ed., pp. 124-146). Newark, DE: International Reading Association.

Valencia, S., & Pearson, P. D. (1987). Reading assessment: Time for a change. *Reading Teacher, 40,* 726-732.

Wiggins, G. (1989). Teaching to the (authentic) test. *Educational Leadership, 46,* 41-47.

Wixson, K. K., Peters, C. W., Weber, E. M., & Roeber, E. D. (1987). New directions in statewide reading assessment. *Reading Teacher, 40,* 749-754.

WQED (Producer). (1988). *First things first* [film]. Pittsburgh: Public Television Outreach Alliance.

Author Note

An earlier version of this report will appear in E. Hiebert (Ed.), *Literacy in a diverse society: Perspectives, practices, and policies*. New York: Teachers College Press, in press.

This page is intentionally blank.